Network Optimized Storage in Leadership Computing Environments

Rockville, MD

February 16th, 2011

Galen M. Shipman

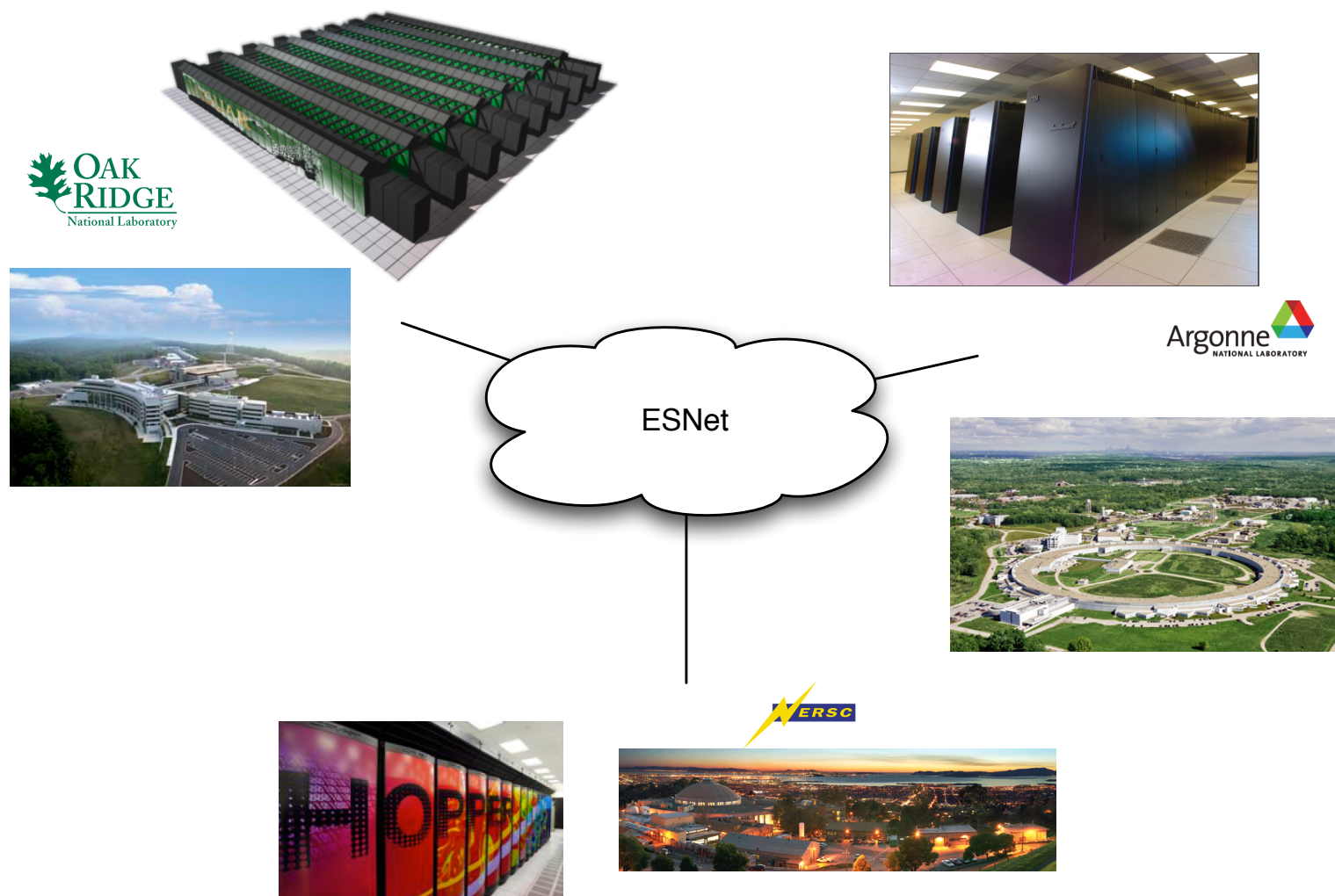**Oak Ridge National Laboratory**

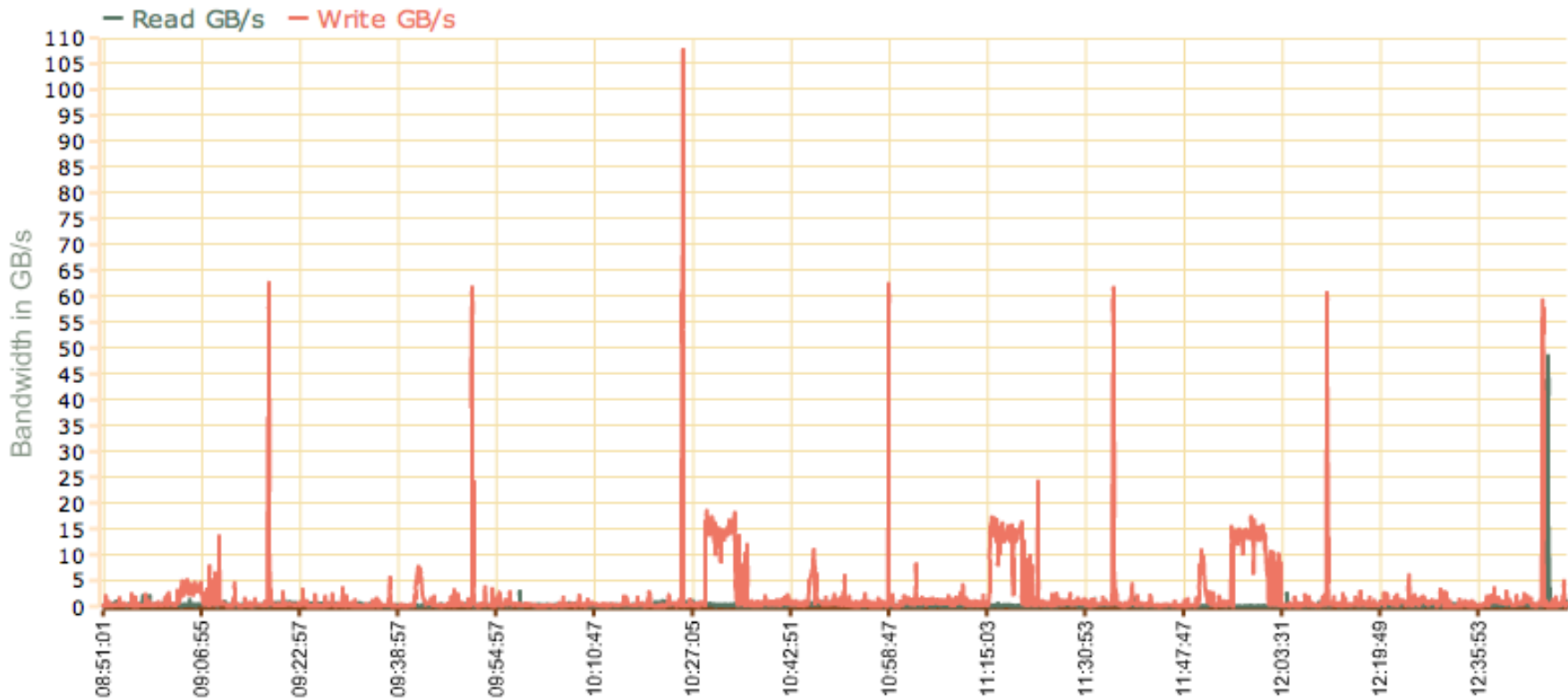# World Class Facilities – Unique Data Management Requirements

ESNet

OLCF facing exponential data growth

Driven by Simulation Platforms
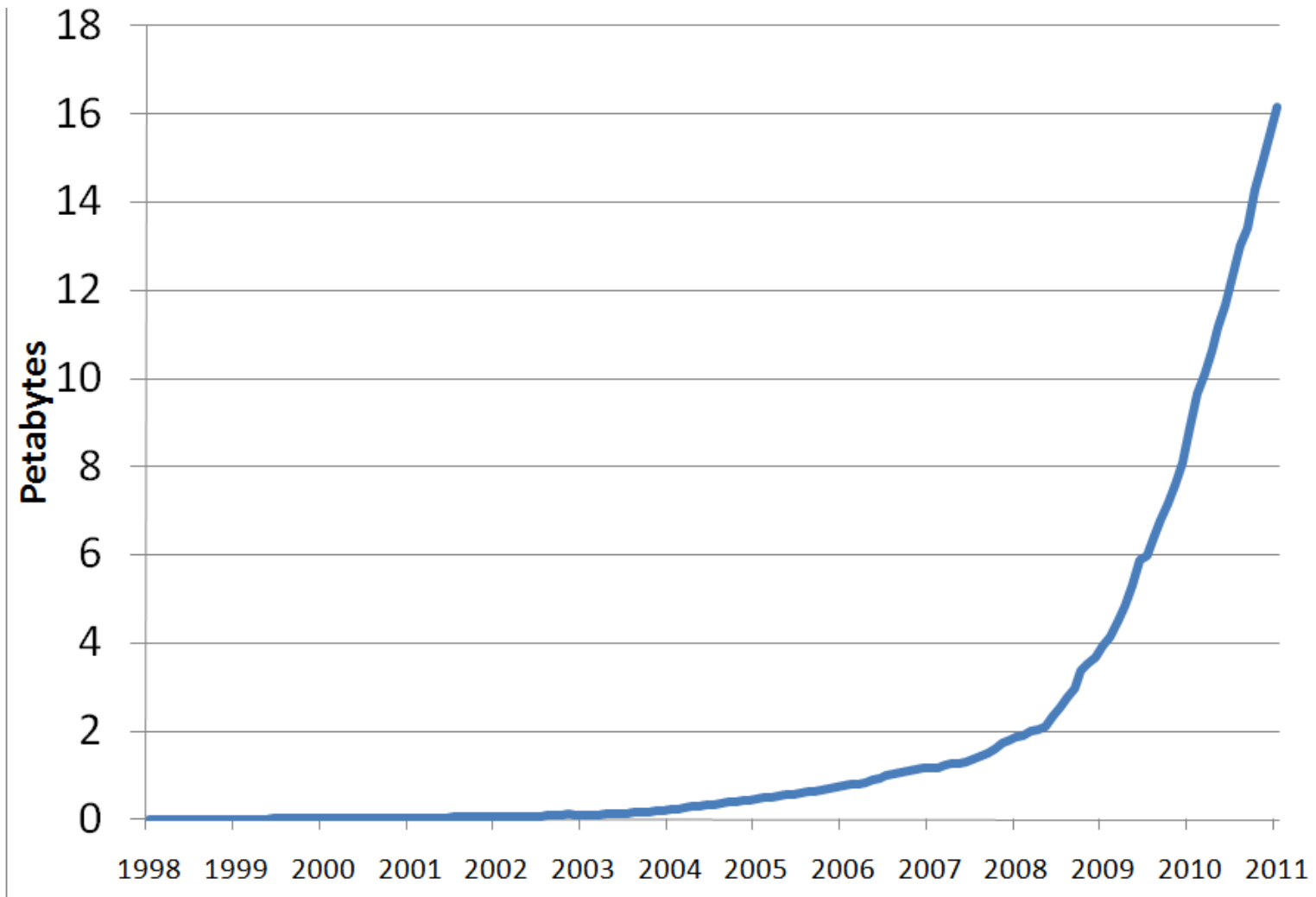
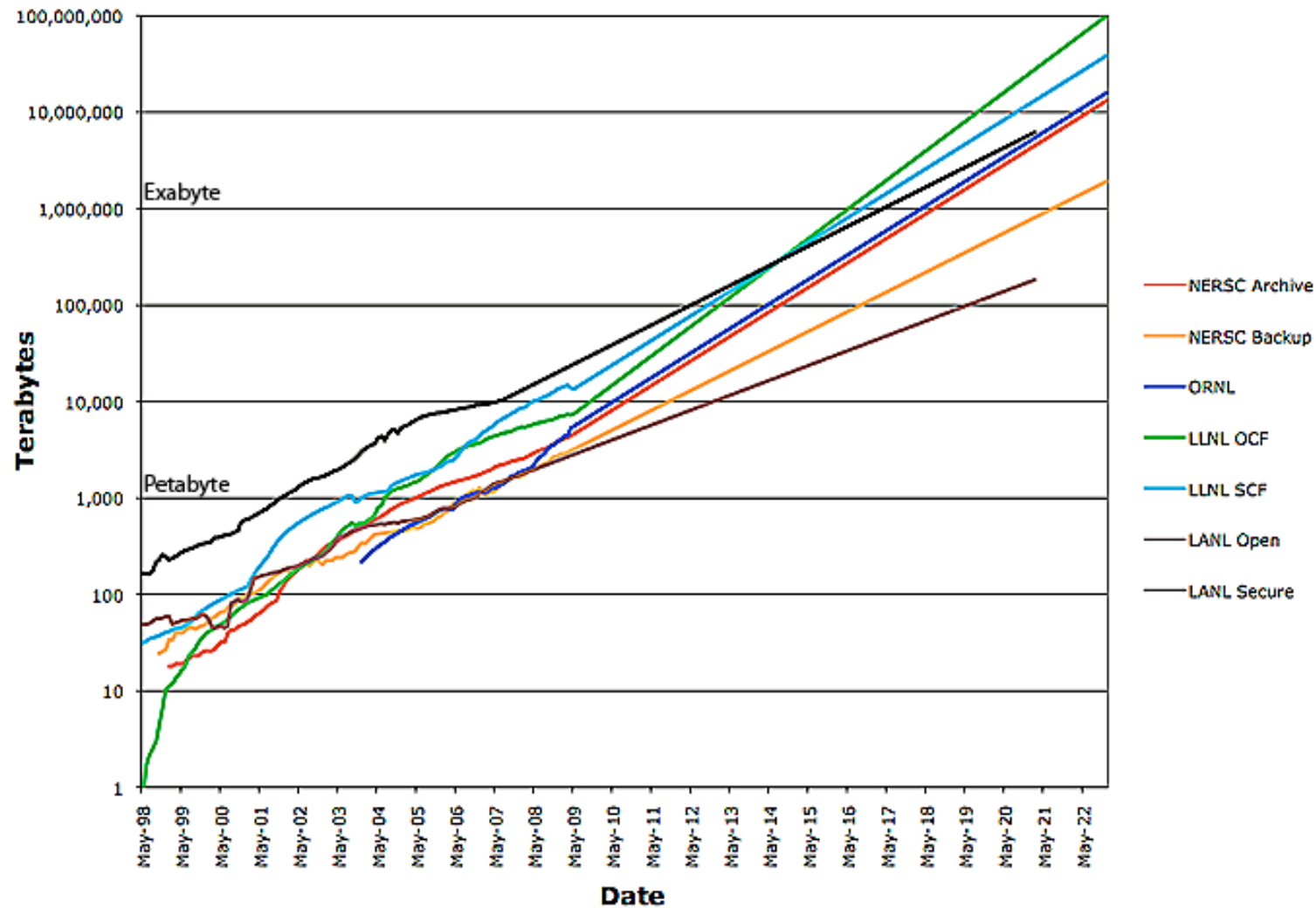**16 PB and growing at more than 30 TB per day.**

Bursty Data Generation Drives I/O Requirements

- **Simulation platform performance cannot be bottlenecked by I/O performance**

Exabyte data in 2018
- ORNL, NERSC, LLNL, LANL

Today we have disparate islands

- **Islands formed around data generators**
  - **Compute platforms**
  - **Scientific instruments**

- **The islands are very far apart**
  - **Data movement between sites is still a challenge**
  - **Users may see 200 MB/sec (~15% utilization)**
    - **Many report 10's of MB/sec**

- **The islands are getting big (at least the OLCF is)**
  - **Exponential growth in both archive and on-line storage**
  - **16.7 Petabytes (24.2 Million files) in HPSS Archival**
  - **4.43 Petabytes (297.4 Million files) in Lustre on-line storage**

- **The islands are littered with small data**
  - **Average file size on spinning disk is 14.8 Megabytes**
  - **Bulk data movement will be limited by Metadata performance**

Current Solutions Are Not Adequate
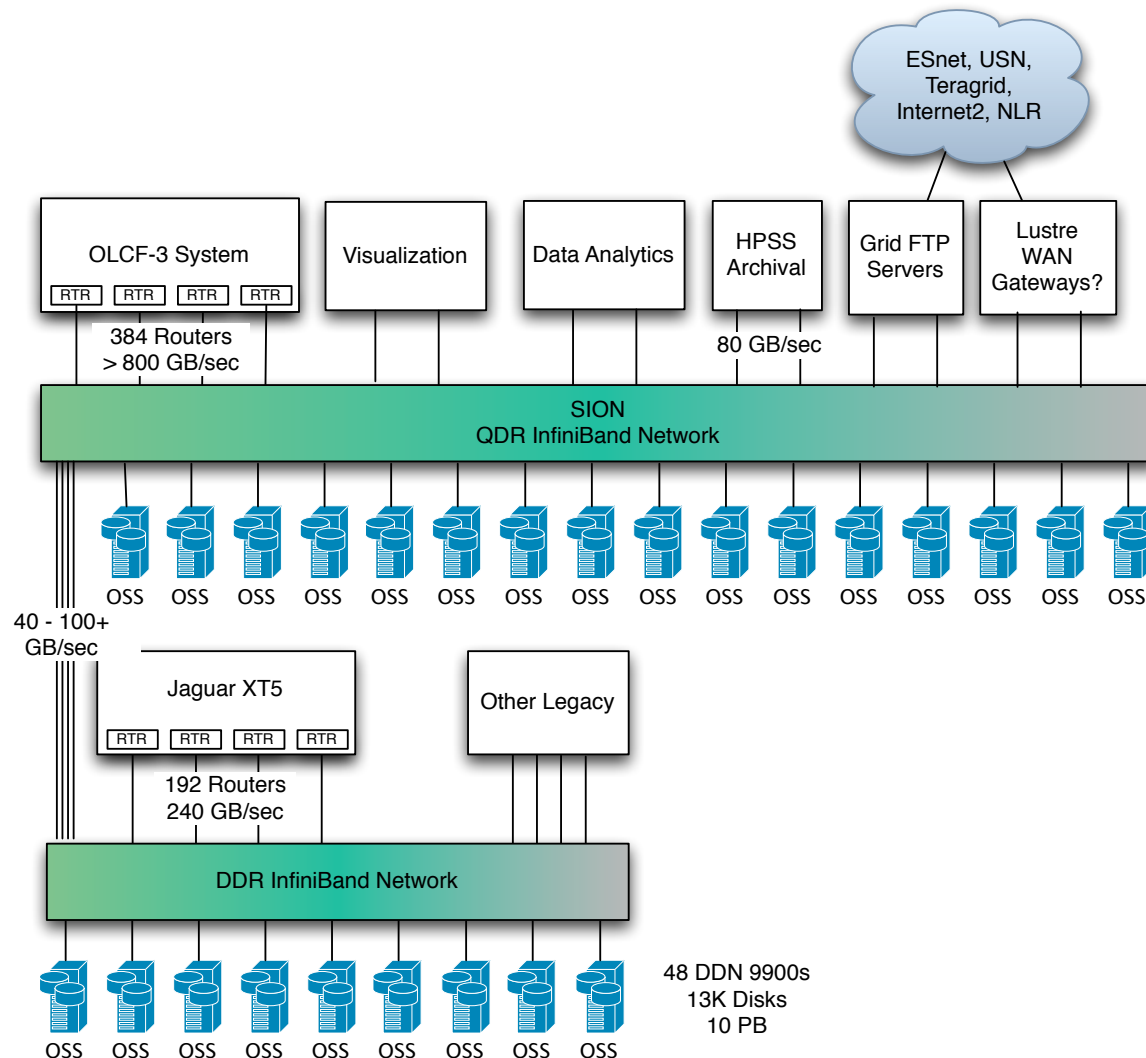
- **Storage systems are optimized for simulation platform, not WAN access**

- **Dedicated data transfer nodes alone won't be sufficient**

- **Optimization at each level of the system software stack is required**
  - WAN/LAN/SAN
  - Data transfer software
  - Storage software
  - Storage hardware

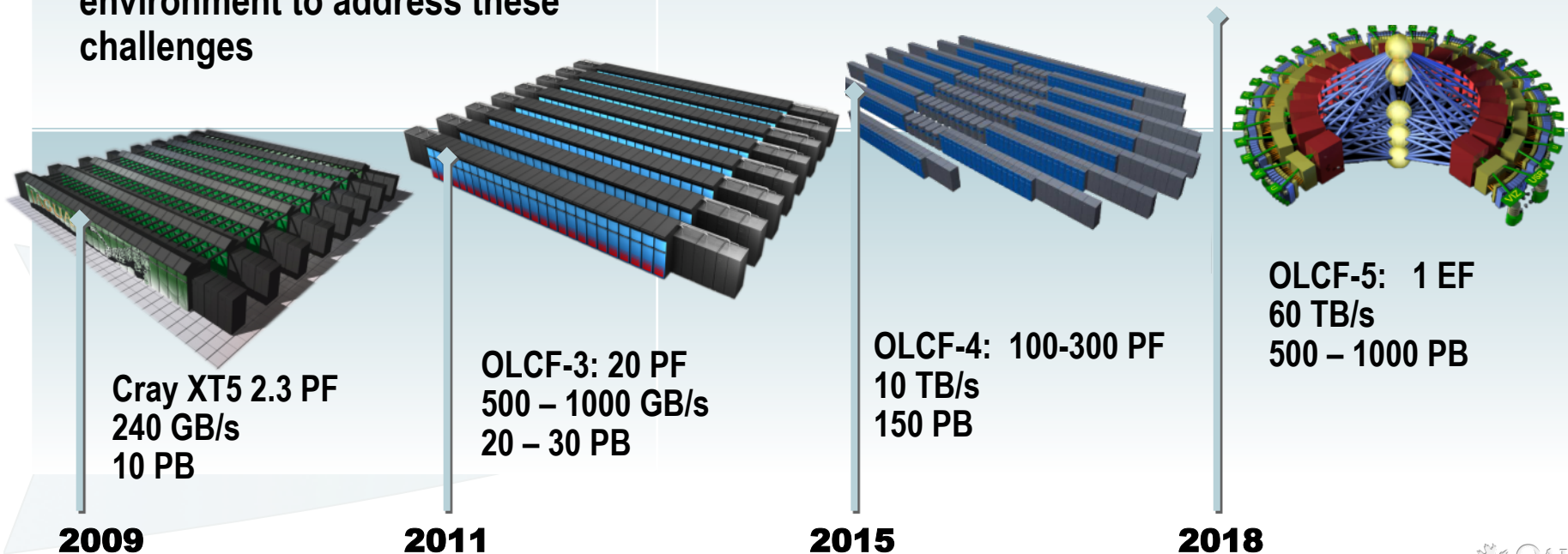- **Commercial solutions are lacking**

Exascale Systems Will Bring New Challenges

| Computational performance (FLOPS) targeted to increase by ~1000x in the next 8 years	Persistent storage performance targeted to increase by ~250x, Capacity increased by ~100x
• Key challenges in power and the ability to leverage dramatic increases in total concurrency • Exascale initiative will drive much of the advances in system architecture and programming environment to address these challenges	• Projected performance improvements pace the rate of improvements delivered over the past 5 years • Key challenges in bandwidth scaling beyond just component count increases

Cray XT5 2.3 PF
240 GB/s
10 PB

OLCF-3: 20 PF
500 – 1000 GB/s
20 – 30 PB

OLCF-4: 100-300 PF
10 TB/s
150 PB

OLCF-5:   1 EF
60 TB/s
500 – 1000 PB

2009   **2011**   **2015**   **2018**

OLCF

OAK RIDGE
National Laboratory

Unfortunately, the world won't get smaller

- **Islands will grow to continents**
  - ¼ Exabyte in 2015
  - over 2 Exabytes in 2018
  - Ariel density gains in magnetic media will help perpetuate this

- **Connectivity improvements are one part of the solution**
  - Will require aggressive deployment of 100Gb (ANI)
  - Terabit capable connectivity will need to be accelerated

- **Bandwidth utilization will be key component of success**
  - End-to-end optimization for data movement
 - Biggest potential gain for end users

OLCF

OAK RIDGE
National Laboratory

An End-to-End Testbed Environment

- **Multi-lab effort spanning ESNet**

- **End-to-end testbed for research, development, QA**
  - Leverage next-gen technologies
  - Allow aggressive transition to Terabit connectivity – End-to-End

- **Integrated R&D**
  - Middleware development
  - Networking stacks
  - Parallel file system
  - Archival storage

- **Specifically targeted to meet the needs of current and future DOE signature facilities**
  - Exascale simulation platforms
  - Instruments
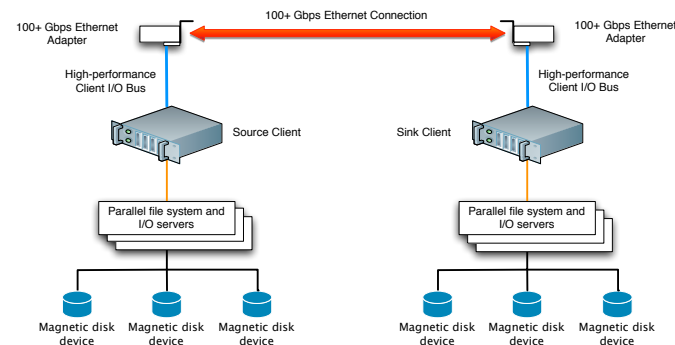
Advanced technologies coupled with integrated development

- **System software**
  - Operating System
  - File System
  - Networking Stack

- **Middleware technologies**
  - Bulk data movers
  - Data management tools
  - Remote analysis
  - Subsetting

- **Advanced HW technologies**
  - SOC
  - CMOS Photonics
  - Interconnect (LAN, SAN, WAN)
  - NVRAM
    - Flash
    - PCM
    - Racetrack
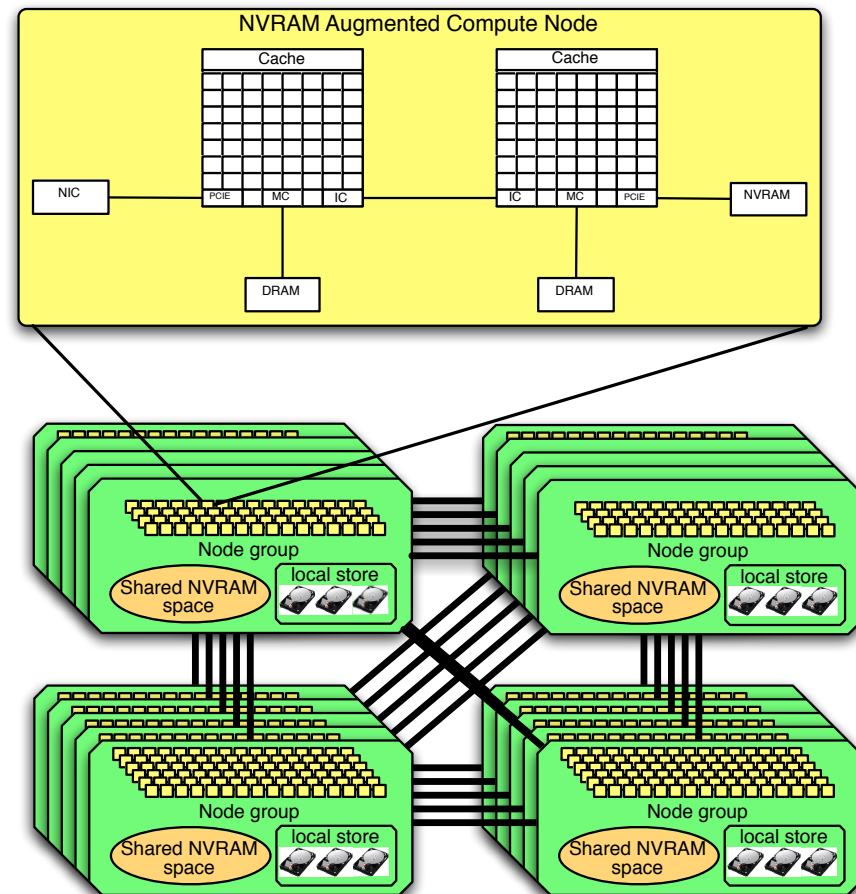
Questions?
Galen Shipman
Email: gshipman@ORNL.Gov

OAK RIDGE
National Laboratory